Ranking WordNet Senses Automatically

CSRP 569

Diana McCarthy and Rob Koeling and Julie Weeds

Department of Informatics, University of Sussex, Falmer, Brighton BN1 9QH

January, 2004

1 Abstract

In word sense disambiguation (WSD), the heuristic of choosing the first listed sense in a dictionary is often hard to beat, especially by systems that do not exploit hand-tagged training data. The problem with using the first sense heuristic, aside from the fact that it does not take surrounding context into account, is that it assumes some quantity of hand-tagged data. Whilst there are some hand-tagged corpora available for some languages, one would expect the frequency distribution of the senses of words to depend on the genre and domain of the text under consideration. For example, one would expect a different predominant sense for *star* if one were looking at scientific astronomy reports compared with popular news. We present work on the use of the WordNet similarity package and a thesaurus automatically acquired from raw textual corpora to rank WordNet noun senses automatically. The results are promising when evaluated against the gold-standard provided by SemCor, giving us a theoretical WSD precision of 60% on an all-words task. Moreover, some of the ranking errors can be explained by differences in the corpus data used to produce the thesaurus compared to this gold-standard. Our experiments also show that the automatic ranking can be used to filter senses which are unseen or infrequent in the gold-standard.

2 Introduction

The first sense heuristic which is often used as a baseline for supervised WSD systems frequently outperforms WSD systems even when they take surrounding context into account. This is shown by the results of the English all-words task in SENSEVAL2 [5] in figure 1 below where the first sense heuristic (labelled 'First Sense 1') was obtained using the frequencies of the sense tagged data provided with WordNet 1.7. [1] The figure

[1] We are indebted to Judita Preiss for this figure.

	File (%)	Genre (%)
Nouns	70	66
Verbs	79	74
Adjectives	25	21

Table 1: Percentages of words with a different predominant sense in SemCor, across files and genres.

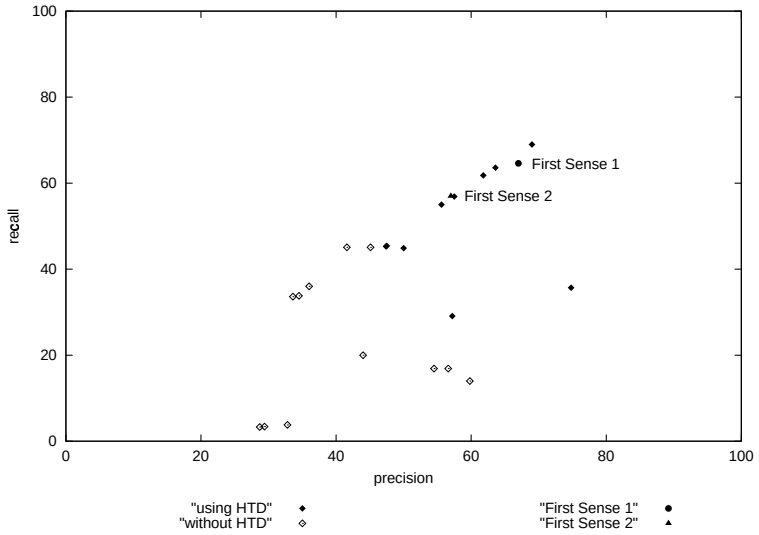

Figure 1: The first sense heuristic compared with SENSEVAL2 results

distinguishes systems which make use of hand-tagged data (using HTD) such as SemCor [16], from those that do not (without HTD). Using the first listed sense in WordNet for the PoS given by the Penn TreeBank would have given a precision and recall of 57% [17], and this is shown with the label 'First Sense 2'. The high performance of the first sense baseline is largely because of the skewed frequency distribution of the senses for most words. Even those systems which show superior performance to this crude heuristic often make use of it where evidence from the context is not sufficient [6]. Whilst a first sense heuristic based on a sense-tagged corpus such as SemCor is clearly useful, there is a strong case for obtaining a first, or predominant, sense from untagged corpus data so that the WSD system can be tuned to the genre or domain at hand. We carried out an analysis of the polysemous nouns, verbs and adjectives in SemCor occurring in more than one SemCor file, and found that a large proportion of nouns and verbs have a different first sense in different files and also in different genres (see table 1). For adjectives there is a lot less variation.

Since many words change their predominant sense depending on the genre or domain and hand-tagging data is a costly process it would be useful to have a method

of ranking senses directly from untagged data. Many WSD systems e.g. [25, 6] use the first sense heuristic within their systems, because it is so powerful. An automatic ranking of senses would be useful for WSD systems, whether or not they also use hand-tagged data for training. Additionally, researchers have used the predominant sense of words to improve lexical acquisition [14, 9] so we believe automatic ranking which could be tuned to the data at hand would be useful for this. As well as being useful for determining the top ranking senses of a word, we hope that a method for ranking senses would also be useful for identifying infrequent and potentially redundant senses.

Assuming that one had a WSD system that could accurately tag a portion of text then one could obtain frequency counts for the senses and rank them with these counts. However, the most accurate WSD systems are those which require manually sense tagged data in the first place, and their accuracy seems to depend on the quantity of training examples [8] available. We are investigating a method of automatically ranking WordNet senses from raw text.

Many researchers are developing automatic thesauruses from automatically parsed data. From inspecting the lists of neighbours of the thesauruses one can see that the ordered neighbours relate to the different senses of the target word in these lists. For example, the neighbours of *star* in a dependency-based thesaurus provided by Lin [2] has the ordered list of neighbours: *superstar, player, teammate, actor* early in the list, but one can also see words that are related to another sense of *star* e.g. *galaxy, sun, world* and *planet* further down the list. The neighbours reflect the various senses of the word to which they relate, *star* in this example. We expect that the quantity and similarity of the neighbours pertaining to different senses will reflect the dominance of the sense to which they pertain. This is because there will be more relational data for the more prevalent senses compared to the less frequent senses. In this paper we describe and evaluate a method for automatically ranking predominant senses of nouns using the neighbours from automatically acquired thesauruses, along with the WordNet Similarity measures [20] for weighting the senses of the target word with the similarity between the neighbours and the senses.

The paper is structured as follows. We discuss our method in the following section and in section 4 describe the experimental setup. Section 5 gives the results of a quantitative evaluation using SemCor as a gold-standard. These results are then discussed in section 6. In section 7 we show further results of applying our method to two domain specific sections of the Reuters corpus for a small sample of words. We describe some related work in section 8 and conclude in section 9.

3 Method

In order to rank senses we first produce automatically acquired thesauruses based on the methods of Lin [12]. We then use the WordNet similarity package [20] which provides an implementation of a host of measures for calculating similarity between words, or senses, within WordNet. To rank the senses of a word (w) we take the k nearest neighbours of that word from the automatic thesaurus and for each neighbour

[2] Downloadable from
http://www.cs.ualberta.ca/~lindek/demos/depsim.htm

we find the WordNet similarity score of the senses of word w to that neighbour, using the sense of the neighbour that maximises this score. We then use the distributionally based similarity scores from the thesaurus for each neighbour along with the WordNet similarity scores between each neighbour and each sense of the target word to rank the senses.

More precisely, let $N_w = \{n_{w1}, n_{w2} \ldots n_{wk}\}$ be the ordered set of the top scoring k neighbours of w from the thesaurus with associated distributional similarity scores $\{dss_{nw1}, dss_{nw2}, \ldots dss_{nwk}\}$. Let WS_w be the set of senses of w. Each neighbour ($n_{wj} \in N_w$) is then associated with a list of the senses ($ws_i \in WS_w$) with associated WordNet similarity scores ($wnss_{ws_i n_{wj}}$) between ws_i and the sense of the n_{wj} that maximises the WordNet similarity. We rank each sense $ws_i \in WS_w$ using:
Ranking Score $ws_i =$

$$\sum_{n_{wj} \in N_w} dss_{n_{wj}} \times \frac{wnss_{ws_i n_{wj}}}{\sum_{ws_{i'} \in WS_w} wnss_{ws_{i'} n_{wj}}} \quad (1)$$

3.1 Acquiring the Automatic Thesaurus

The thesaurus was acquired using the method described by Lin [12]. For input to the thesaurus we used grammatical relation data extracted from the the 90 million words of written English from the British National Corpus (BNC) using an automatic parser [3]. For each noun we considered the co-occurring verbs in the direct object and subject relation, the modifying nouns in noun-noun relations and the modifying adjectives in adjective-noun relations. A noun, n, is thus described by a set of co-occurrence triples $< n, r, w >$ and associated frequencies, where r is a grammatical relation and w is a possible co-occurrence with n in that relation. For every pair of nouns, where each noun had a total frequency in the triple data of 10 or more, we computed their distributional similarity using the measure given by Lin [12]. If $T(n)$ is the set of co-occurrence types (r, w) such that $I(n, r, w)$ is positive then the similarity between two nouns, n_1 and n_2, can be computed as:

$$\frac{\sum_{(r,w) \in T(n_1) \cap T(n_2)} (I(n_1, r, w) + I(n_2, r, w))}{\sum_{(r,w) \in T(n_1)} I(n_1, r, w) + \sum_{(r,w) \in T(n_2)} I(n_2, r, w)}$$

where:

$$I(n, r, w) = \log \frac{P(w|n \cap r)}{P(w|r)}$$

A thesaurus entry of size k for a target noun n can then be defined as the k most similar nouns to noun n.

3.2 The WordNet Similarity Package

We use the WordNet Similarity Package 0.05 and WordNet version 1.6. The WordNet Similarity package supports a range of WordNet similarity scores. We experimented

with 6 of these [3] and briefly summarise them below, for a more detailed summary see [19]. The measures provide a similarity score between two WordNet senses ($c1$ and $c2$).

lesk [2] This score maximises the number of overlapping words in the gloss, or definition, of the senses. It uses the glosses of semantically related (according to WordNet) senses too.

lch [10] $lch(c1, c2) = max[-log(\frac{SL(c1,c2)}{2 \times D})]$ where $SL(c1, c2)$ is the length of the path with the minimum number of intervening classes between $c1$ and $c2$ and D is the maximum depth of the hierarchy, i.e. the largest distance between any leaf and the root of the hierarchy.

edge A simplistic measure which inverts the edge counts between c1 and c2.
$edge(c1, c2) = \frac{1}{SL(c1,c2)}$

res [21] This score uses frequency counts to populate the hierarchy. It is calculated as the information content (IC) of the class ($c3$) which is the lowest possible subsumer of the two senses.
$res(c1, c2) = IC(c3)$ where $IC(c3) = -log(p(c3))$

jcn [7] Actually a distance measure which relates to the res measure, but with a notion of path-length:
$dist_{jcn}(c1, c2) = IC(c1) + IC(c2) - 2 \times res(c1, c2)$
This is transformed from a distance measure in the WN-Similarity package by taking the reciprocal:
$jcn(c1, c2) = 1/dist_{jcn}(c1, c2)$

lin [13]. The similarity between two classes is the ratio of the information needed to state the shared information and the amount of information to describe them independently. This is related to the jcn measure.
$lin(c1, c2) = \frac{2 \times res(c1,c2)}{IC(c1)+IC(c2)}$

4 Experimental Setup

In order to evaluate our rankings we use the data in SemCor as a gold-standard. This is not ideal since we expect that the rankings within SemCor will be different to those that would be obtained from the BNC, from which we obtain our thesaurus. Nevertheless, since many systems performed well on the English all-words task for SENSEVAL2 using the frequency information in SemCor we believe this is a reasonable approach for evaluation.

We generated a thesaurus entry for all polysemous nouns which occurred in SemCor with a frequency greater than 2, and in the BNC with a frequency of at least 10 in the grammatical relations listed in section 3.1 above. Three of the measures (jcn, lin,

[3] There are a further two scores in this package, but the implementation in this version of the package was too slow for us to complete the experiments with them. These other measures did not perform as well as lesk and jcn in previous WSD experiments [19]

res) in the WordNet similarity package are produced using corpus data to obtain the IC of a class. We evaluated the rankings using four variations in obtaining the counts for these with data from (i) the BNC corpus and a resnik count option which is implemented in the WordNet similarity package, (hereafter referred to as rc) (ii) the Brown corpus with rc (iii) the Brown corpus and (iv) the default counts derived from SemCor. All the results shown here are those with the size of thesaurus entries (k) set to 50. We repeated the experiment with the BNC data for jcn using $k = 10, 30, 50$ and 70 however, the number of neighbours used gave only minimal changes to the results so we do not report them here.

4.1 Evaluation Metrics

For evaluation we compare the automatic rankings with data in SemCor. We calculate the accuracy of finding the predominant sense, when there is indeed one sense with a higher frequency than the others for this word in SemCor (PS_{acc}). We also use a pairwise agreement (PWA) of rankings devised by Briscoe and Carroll [4]. This takes each pair of senses at positions (x, y) in the automatic ranking such that $x < y$ and calculates the percentage agreement of all such pairs where the pair is ordered the same as the gold-standard ranking. We also calculate the WSD accuracy that would be obtained on SemCor, when using our top ranked sense in all contexts (WSD_{sc}).

As well as being useful for determining the top ranking senses of a word, we hope that our method will be good for identifying infrequent and potentially redundant senses. This would be particularly useful when applying the method to domain specific text, rather than balanced text like the BNC. To evaluate this we take a threshold of the ranking score which would filter a constant percentage ($F\%$) of the sense types in our experiment. We then calculate the percentage of these sense types that do not occur in SemCor at all ($Ftype_{acc}$), and for those that do, we calculate $Ftok_{err}$ which is the percentage of sense tokens that would be filtered incorrectly using this ranking score as a threshold (i) from the entire set of nouns, and (ii) from the subset of nouns that have at least one sense filtered.

5 Results

The results in table 2 show the accuracy of the ranking with respect to SemCor over the entire set of 2595 polysemous nouns in SemCor with the 6 similarity measures. The results in this table are obtained using the BNC rc IC file for the res, jcn and lin measures. The random baseline for choosing the predominant sense over all these words ($\sum_{w \in Words} \frac{1}{|senses_w|}$) is 32%. We see that all WordNet similarity measures beat this baseline. The random baseline for WSD_{sc}, obtained by dividing each word token from SemCor for this set of words by the number of senses of that word, is 24%. Again, the automatic ranking outperforms this by a large margin.

Table 3 gives results for the sense ranking for the 2595 nouns given the four variations for obtaining the IC data described in section 4 above. The ranking results do not differ substantially except that we do get a significant increase in performance using the default IC files provided with the WordNet similarity package. This is particularly

measure	PS_{acc} %	PWA %	WSD_{sc} %
lesk	54	48	48
lch	49	48	43
edge	50	49	44
res	48	45	39
jcn	54	50	46
lin	50	46	43

Table 2: Results with the BNC thesaurus for a range of WordNet similarity scores

BNC resnik count			
measure	PS_{acc} %	PWA %	WSD_{sc} %
res	48	45	39
jcn	54	50	46
lin	50	46	43
Brown resnik count			
res	47	45	39
jcn	55	50	46
lin	50	47	43
Brown			
res	47	45	39
jcn	54	50	46
lin	50	46	42
default (SemCor)			
res	47	45	37
jcn	69	68	55
lin	62	64	49

Table 3: Results with the BNC thesaurus for different sources of IC

the case with the jcn and lin metrics which use the IC of the senses of the words directly. These default files rely on the sense tagged data from SemCor. To avoid using sense-tagged data (and in particular our test set) we revert to using the BNC rc IC files for the rest of the work described in this paper.

The lesk and jcn measures show the best performance on the scores that evaluate the rankings. The lesk measure was considerably slower than the measures, such as jcn, which rely on the precompiled files from the corpus data. Since all measures give comparable results we restricted our remaining experiments to jcn because this gave good results for the sense ranking metrics, and is efficient, given the precompilation of the IC files.

Table 4 displays the results obtained when varying the filter threshold (F%), giving the number of sense types filtered from the full set of 10687 sense types for all 2595 polysemous nouns. $Ftype_{acc}$, described above in section 4.1, is the percentage of these

F%	# Ftypes	$Ftype_{acc}$	$Ftok_{err_i}$	$Ftok_{err_{ii}}$
1	99	57	0.04	25
3	298	57	1.3	33
5	508	57	2.1	32
10	998	56	5.3	44

Table 4: Filtering results

types that do not occur at all in SemCor. $Ftok_{err_i}$ is the percentage of tokens filtered from SemCor in error using this threshold, and $Ftok_{err_{ii}}$ is that percentage for the subset of nouns which have at least one sense filtered using that threshold.

The results show that the majority of sense types filtered are those that do not occur in SemCor. The baseline for this task is 38% since that is the percentage of sense types for the set of polysemous nouns that do not occur in SemCor. Interestingly, increasing the threshold seems to filter a higher percentage of tokens that should not be removed, but the percentage of sense types that are filtered correctly remains similar. The filtering threshold using a percentage of sense types results in some words having many more senses filtered than others. In the future we plan to investigate this more carefully and determine if a word specific threshold would be more appropriate.

6 Discussion

From manual analysis, there are cases where the automatic ranking is at odds with SemCor, yet the automatic ranking is intuitively plausible. This is to be expected regardless of any inherent shortcomings of the ranking technique since the senses within SemCor will differ compare to those of the BNC. For example, in WordNet the first listed sense of *pipe* is **tobacco pipe**, and this is ranked joint first according to the Brown files in SemCor with the second sense **tube made of metal or plastic used to carry water, oil or gas etc....** The automatic ranking from the BNC data lists the latter **tube** sense first. This seems quite reasonable given the nearest neighbours: [4] *tube, cable, wire, tank, hole, cylinder, fitting, tap, cistern, plate...*

Since SemCor is derived from the Brown corpus, which predates the BNC by 30 years [5] and contains a higher proportion of fiction [6], the high ranking for the **tobacco pipe** sense according to SemCor seems quite plausible. It could however be that this example highlights a problem with using automatically acquired thesauruses for some cases. It may be that the **tobacco pipe** sense is simply demoted because it does not occur in a wide variety of contexts and so it is not adequately reflected in the list of neighbours.

[4] We show the first 10 for the sake of brevity.

[5] The text in the Brown corpus was produced in 1961, whereas the bulk of the written portion of the BNC contains texts produced between 1975 and 1993.

[6] 6 out of the 15 Brown genres are fiction, including one specifically dedicated to detective fiction, whilst only 20% of the BNC text represents imaginative writing, the remaining 80% being classified as informative.

Another example where the ranking is intuitive, is *soil*. The first ranked sense according to SemCor is the **filth, stain: state of being unclean** sense whereas the automatic ranking lists **dirt, ground, earth** as the first sense, which is the second ranked sense according to SemCor. This seems intuitive given our expected relative usage of these senses in modern British English, however, we have not manually hand-tagged the BNC data to verify this.

Even given the difference in text of SemCor and the BNC the results are encouraging, especially given that the WSD performance cited here is for polysemous nouns. In the English all-words SENSEVAL2, 25% of the noun data was monosemous. Thus, if we used the sense ranking as a heuristic for an "all nouns" task we would expect to get precision in the region of 60% which would have outperformed all systems not using hand-tagged data on the all words task, and many of those that did. We believe that our technique may be of benefit to lexical acquisition systems requiring wide coverage WSD, or to high precision WSD systems wishing to make use of a first sense heuristic. Of course, selecting the first sense from SemCor outperforms our automatic ranking when evaluated on that same gold-standard. The question then is whether automatic ranking works better than a SemCor derived ranking when turning to a new text type.

7 Experiments with Reuters data

We are interested to see if our method is able to capture the change in ranking of senses for documents from different domains. In order to do that we applied our method to two specific sections of the Reuters corpus.

7.1 Reuters Corpus

The Reuters corpus [23] is a collection of about 810,000 Reuters, English Language News stories (covering the period August 1996 to August 1997). Many of the news stories are economy related, but several other topics are included too. We have selected documents from the SPORTS domain (topic code: GSPO) and a limited number of documents from the FINANCE domain (topic codes: ECAT (ECONOMICS) and MCAT (MARKETS)).

The SPORT corpus consists of 35317 documents (about 9.1 million words). The FINANCE corpus consists of 117734 documents (about 32.5 million words). We acquired thesauruses for these corpora using the procedure described in section 3.1.

7.2 The Experiment

For this experiment there is no existing gold-standard that we could use for a quantitative evaluation. We therefore decided to select a limited number of words (11) and to evaluate these words manually. The words included in this experiment are not a random sample, since we anticipated different predominant senses in the SPORT and FINANCE domain for these words[7]. We are planning to carry out a more extensive

[7]There were 7 more words included in this experiment. Those words proved to be neither good nor bad examples. They were not particularly illustrative examples and therefore excluded from discussion here.

Word	PS BNC	PS FINANCE	PS SPORT
pass	1 (**accomplishment**)	14 (**attempt**)	15 (**throw**)
share	2 (**portion, asset**)	2	2
division	4 (**admin. unit**)	4	6 (**league**)
head	1 (**body part**)	4 (**leader**)	4
loss	2 (**transf. property**)	2	8 (**death, departure**)
competition	2 (**contest, social event**)	3 (**rivalry**)	2
match	2 (**contest**)	7 (**equal, person**)	2
tie	1 (**neckwear**)	2 (**affiliation**)	3 (**draw**)
strike	1 (**work stoppage**)	1	6 (**hit,success**)
striker	1 (**athlete**)	2 (**sailor**)	1
goal	1 (**end,mental object**)	1	2 (**score**)

Table 5: Domain specific results

evaluation in the near future.

7.3 Discussion

The results for the experiments are summarized in table 5. The results are promising. Most words show the change in predominant sense (PS) that we anticipated. It is not always intuitively clear which of the senses to expect as predominant sense for either a particular domain or for the BNC, but the first senses of words like *division* and *goal* shift towards the more specific senses (**league** and **score** respectively). Moreover, the chosen senses of the word *tie* proved to be a textbook example of the behaviour we expected.

The word *share* is among the words whose predominant sense remained the same for all three corpora. We anticipated that the **stock certificate** sense would be chosen for the FINANCE domain, but this did not happen. However, that particular sense ended up higher in the ranking for the FINANCE domain. The word *striker* is also an interesting case where the ranking is somewhat counter-intuitive. We expected the first sense for the FINANCE domain to be **nonworker**. The thesaurus, however, gave words such as: *pilot*, *trucker*, *driver*, *miner*, *farmer*, *teacher*, *nurse*, *steelworker*, etc... as nearest neighbours, since these occur in similar contexts. Considering these neighbours, the chosen sense **sailor** is to be expected.

This experiment shows the behaviour we were expecting. However, we need to evaluate this quantitatively. We are planning an extensive quantitative evaluation in the near future.

The fact that we can use any collection of raw text for our method, means that we have some excellent opportunities to use the web as a corpus. There are several directories available on the web with pointers to webpages about certain topics. We could harvest texts from the web using these directories and create corpora for specific domains automatically. These domain specific corpora could then be used for sense ranking as described in this paper and also for training a text classifier which could be

used both to obtain further input data for domain specific sense ranking and for use in determining the domain for application of the domain specific ranking.

8 Related Work

To our knowledge there is no other work on automatically ranking senses. Of course this could be done by using an unsupervised WSD system to tag text and taking the resulting rankings. The major problem with this is that the accuracy of unsupervised systems does not seem to be sufficient. The answers for system sussex-sel [15] [8] would give a PS_{acc} of 32% for finding the first sense according to WordNet 1.7. Systems that use training data, such as SemCor, would undoubtedly do better on ranking, but it would probably be better to use the training data directly.

Patwardhan et. al. [19] have used the WordNet similarity packages for WSD, and evaluated on the SENSEVAL2 English lexical sample data. The results look comparable to others that do not make use of hand-tagged data, with the optimum accuracy at 39%. Interestingly, variation of the IC files did not affect their results much, as with ours, however, unlike our results this was also the case where the sense-tagged data in the SemCor files was used. The task is different though, in that we are evaluating rankings and not performing WSD. Additionally, our gold-standard was the same as that used for the default IC files, whereas they used the SemCor frequency counts, but then applied their WSD to the SENSEVAL2 lexical sample data.

There has been some related work on using automatic thesauruses for discovering word senses from corpora. Pantel and Lin [18]. In this work the lists of neighbours are themselves clustered to bring out the various senses of the word. They evaluate using the lin measure described above in section 3.2 to determine the precision and recall of these discovered classes with respect to WordNet synsets. This method obtains precision of 61% and recall 51%. If WordNet sense distinctions are not ultimately required then discovering the senses directly from the neighbours list is useful because sense distinctions discovered are relevant to the corpus data and new senses can be found. In contrast, we use the neighbours lists and WordNet similarity measures to rank and filter WordNet senses. We believe automatic ranking and filtering techniques will be useful for systems that rely on WordNet, for example those that use it for lexical acquisition or WSD.

9 Conclusions

We have devised a method that uses raw corpus data to automatically rank the senses of nouns in WordNet. We use an automatically acquired thesaurus and a WordNet Similarity measure – such as those available in the WordNet similarity package. The ranking results are promising when evaluated quantitatively on SemCor, giving us a theoretical WSD precision of 60% on an all-words task. In many cases the sense ranking provided in Semcor differs to that obtained because of differences of the input data from

[8] This system obtained the highest precision on the all words task for those systems that did not make use of hand-labelled data such as SemCor.

which the automatic thesaurus was produced. In the future, we hope to quantitatively evaluate the rankings for domain specific corpora. In particular we plan to use the rankings to demonstrate an improvement in lexical acquisition, and also examine the effects of filtering low ranking senses prior to lexical acquisition. We hope also to use the difference in rankings between balanced corpora and domain specific ones to isolate words having very different neighbours, and therefore rankings, in the different corpora. As regards the filtering, it may be that we need a word specific threshold before filtering is applied and we plan to do some experiments in this direction.

There is plenty of scope for further work. WordNet is very fine-grained. From manual analysis the thesaurus method often picks a closely related sense to the gold-standard, or anticipated, predominant sense. We hope to look at automatic methods for clustering WordNet related senses, such as those proposed by Agirre and Lopez de Lacalle [1]. We have not yet performed any analysis on the categories of nouns for which our ranking method works best. Such analysis would also be useful, though we suspect that performance depends on the granularity and that the method will work best on those nouns with clear distinctions.

We also want to investigate whether the frequency of the nouns has a bearing on performance since Lin's measure of distributional similarity has been shown to perform poorly on semantic tasks for low frequency words [24]. It would be worth exploring other distributional similarity measures for producing the thesaurus such as alpha-skew divergence [11] and those proposed by Weeds and Weir. Additionally, we need to determine whether senses which do not occur in a wide variety of contexts fare badly using distributional measures of similarity, and what can be done to combat this problem.

To date we have only used this method on nouns. We hope to try it out on verbs, but we will first need to look into the performance of WordNet similarity measures for verbs. Resnik and Diab [22] discuss the fact that verb similarity is a different to noun similarity because of the different properties of verbs, such as event structure. They found lower inter-rater agreement for humans judging the similarity of verbs than has been obtained in similar experiments for nouns.

The lesk and jcn measures performed the best in our evaluations using SemCor as a gold-standard. Since both of these measures use different types of information we could try using a combination of similarity scores within our ranking score.

Acknowledgements

We would like to thank Siddharth Patwardhan and Ted Pedersen for making the WN Similarity package publically available.

References

1. Eneko Agirre and Oier Lopez de Lacalle, *Clustering wordnet word senses*, Recent Advances in Natural Language Processing (Borovets, Bulgaria).

2. Satanjeev Banerjee and Ted Pedersen, *An adapted Lesk algorithm for word sense disambiguation using WordNet*, Proceedings of the Third International Confer-

ence on Intelligent Text Processing and Computational Linguistics (CICLING-02) (Mexico City).

3. Edward Briscoe and John Carroll, *Robust accurate statistical annotation of general text*, Proceedings of the Third International Conference on Language Resources and Evaluation (LREC) (Las Palmas, Canary Islands, Spain), pp. 1499–1504.

4. Ted Briscoe and John Carroll, *Automatic extraction of subcategorization from corpora*, Proceedings of the Fifth Applied Natural Language Processing Conference, pp. 356–363.

5. Scott Cotton, Phil Edmonds, Adam Kilgarriff, and Martha Palmer, SENSEVAL-2, http://www.sle.sharp.co.uk/senseval2/, 1998.

6. Véronique Hoste, Anne Kool, and Walter Daelemans, *Classifier optimization and combination in the English all words task*, Proceedings of the SENSEVAL-2 workshop, pp. 84–86.

7. Jay Jiang and David Conrath, *Semantic similarity based on corpus statistics and lexical taxonomy*, International Conference on Research in Computational Linguistics (Taiwan).

8. Adam Kilgarriff and Martha Palmer, *Introduction to the special issue on SENSEVAL*, Computers and the Humanities. SENSEVAL Special Issue **34** (2000), no. 1–2, 1–13.

9. Anna Korhonen, *Subcategorization acquisition*, Ph.D. thesis, University of Cambridge, 2002.

10. Claudia Leacock and Martin Chodorow, *Combining local context and WordNet similarity for word sense disambiguation*, WordNet: an Electronic Lexical Database (Christiane Fellbaum, ed.), MIT Press, 1998, pp. 268–283.

11. Lillian Lee, *Measures of distributional similarity*, Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pp. 25–32.

12. Dekang Lin, *Automatic retrieval and clustering of similar words*, Proceedings of COLING-ACL 98 (Montreal, Canada).

13. _____, *An information-theoretic definition of similarity*, Proceedings of the 15th International Conference on Machine Learning (Madison, WI).

14. Diana McCarthy, *Word sense disambiguation for acquisition of selectional preferences*, Proceedings of the ACL/EACL 97 Workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, pp. 52–61.

15. Diana McCarthy, John Carroll, and Judita Preiss, *Disambiguating noun and verb senses using automatically acquired selectional preferences*, Proceedings of the SENSEVAL-2 workshop, pp. 119–122.

16. A. Miller, George, Claudia Leacock, Randee Tengi, and Ross T Bunker, *A semantic concordance*, Proceedings of the ARPA Workshop on Human Language Technology, Morgan Kaufman, pp. 303–308.

17. Martha Palmer, Christiane Fellbaum, Scott Cotton, Lauren Delfs, and Hoa Trang Dand, *English tasks: All-words and verb lexical sample*, Proceedings of the SENSEVAL-2 workshop, pp. 21–24.

18. Patrick Pantel and Dekang Lin, *Discovering word senses from text*, Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining (Edmonton, Canada), pp. 613–619.

19. Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen, *Using measures of semantic relatedness for word sense disambiguation*, Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics (Mexico City).

20. Siddharth Patwardhan and Ted Pedersen, *The cpan wordnet::similarity package*, http://search.cpan.org/author/SID/WordNet-Similarity-0.03/, 2003.

21. Philip Resnik, *Using information content to evaluate semantic similarity in a taxonomy*, 14th International Joint Conference on Artificial Intelligence (Montreal).

22. Philip Resnik and Mona Diab, *Measuring verb similarity*, Proceedings of the 22nd Annual Meeting of the Cognitive Science Society (CogSci 2000), pp. 399–404.

23. Tony G. Rose, Mary Stevenson, and Miles Whitehead, *The reuters corpus volume 1 - from yesterday's news to tomorrow's language resources*, Proc. of Third International Conference on Language Resorces and Evaluation (Las Palmas de Gran Canaria).

24. Julie Weeds and David Weir, *A general framework for distributional similarity*, Proceedings of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.

25. Yorick Wilks and Mark Stevenson, *The grammar of sense: using part-of speech tags as a first step in semantic disambiguation*, Natural Language Engineering **4** (1998), no. 2, 135–143.